ISBA World Meeting, Forte Village Sardinia, Italy, 13 – 17 June 2016

# Priors for frequentists, consistency beyond Schwartz

Bas Kleijn, KdV Institute for Mathematics

Universiteit van Amsterdam

# Part I

## Introduction

# Bayesian and Frequentist statistics

| | | |
|---|---|---|
| sample space | $(\mathscr{X}, \mathscr{B})$ | measurable space |
| *i.i.d.* data | $X^n = (X_1, \ldots, X_n) \in \mathscr{X}^n$ | frequentist/Bayesian |
| model | $(\mathscr{P}, \mathscr{G})$ | model subsets $B, V \in \mathscr{G}$ |
| prior | $\Pi : \mathscr{G} \to [0, 1]$ | probability measure |
| posterior | $\Pi(\,\cdot\,\| X^n) : \mathscr{G} \to [0, 1]$ | Bayes's rule, inference |

| | | |
|---|---|---|
| Frequentist | assume there is $P_0$ | $X^n \sim P_0^n$ |
| Bayes | assume $P \sim \Pi$ | $X^n \,\|\, P \sim P^n$ |

# Definition of the posterior

**Definition 4.1** *Assume that all $P \mapsto P^n(A)$ are $\mathscr{G}$-measurable. Given prior $\Pi$, a posterior is any $\Pi(\,\cdot\,|X^n = \cdot) : \mathscr{G} \times \mathscr{X}^n \to [0,1]$*

*(i) For any $G \in \mathscr{G}$, $x^n \mapsto \Pi(G|X^n = x^n)$ is $\mathscr{B}^n$-measurable*

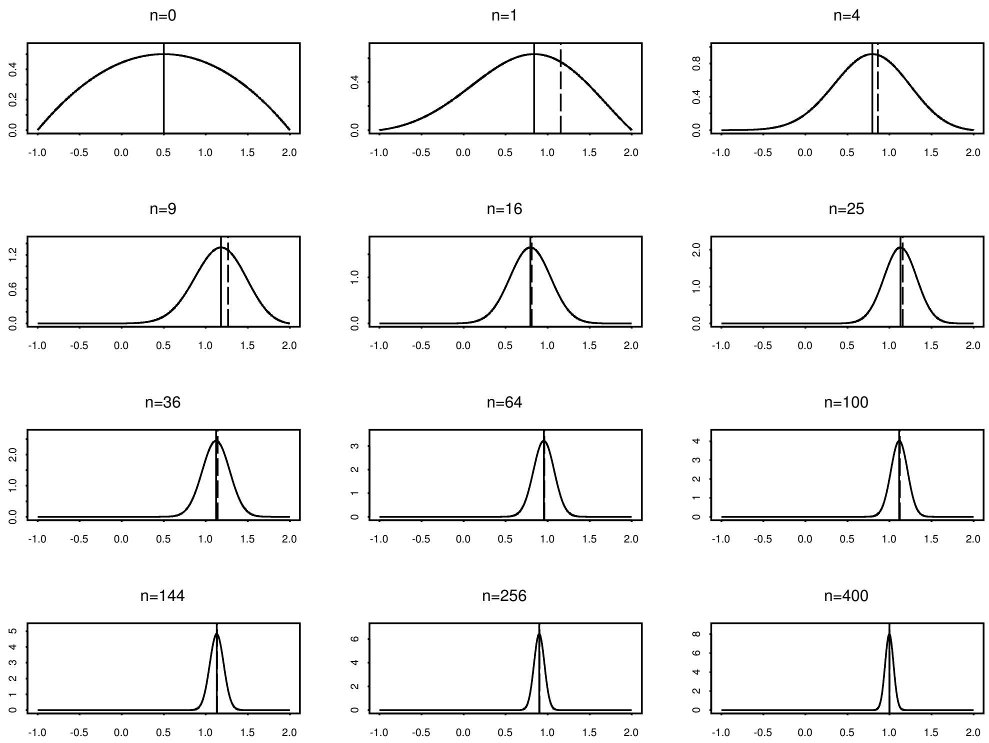
*(ii) (Disintegration) For all $A \in \mathscr{B}^n$ and $G \in \mathscr{G}$*

$$\int_A \Pi(G|X^n)\, dP_n^{\Pi} = \int_G P^n(A)\, d\Pi(P)$$

*where $P_n^{\Pi} = \int P^n\, d\Pi(P)$ is the prior predictive distribution*

**Remark 4.2** *For frequentists $(X_1, \ldots, X_n) \sim P_0^n$, so assume $P_0^n \ll P_n^{\Pi}$*

# Asymptotic consistency of the posterior




**Definition 5.1** *Given a model $\mathscr{P}$ with topology and a Borel prior $\Pi$, the posterior is consistent at $P \in \mathscr{P}$ if for every open nbd $U$ of $P$*

$$\Pi(U|X^n) \xrightarrow{P} 1$$

# Doob's and Schwartz's consistency theorems

**Theorem 6.1** *(Doob (1948))*
*Let $\mathscr{P}$ and $\mathscr{X}$ be Polish spaces and let* $\Pi$ *be a Borel prior. Assume that $P \mapsto P^n(A)$ is Borel measurable for all $n, A$. Then the posterior is consistent at $P$, for* $\Pi$*-almost-all $P \in \mathscr{P}$*

**Remark 6.2** *(Schwartz (1961), Freedman (1963)) Not frequentist!*

**Theorem 6.3** *(Schwartz (1965))*
*Let $X_1, X_2, \ldots$ be an i.i.d.-sample from $P_0 \in \mathscr{P}$. Let $\mathscr{P}$ be Hellinger totally bounded and let* $\Pi$ *be a Kullback-Leibler (KL-)prior, i.e.*

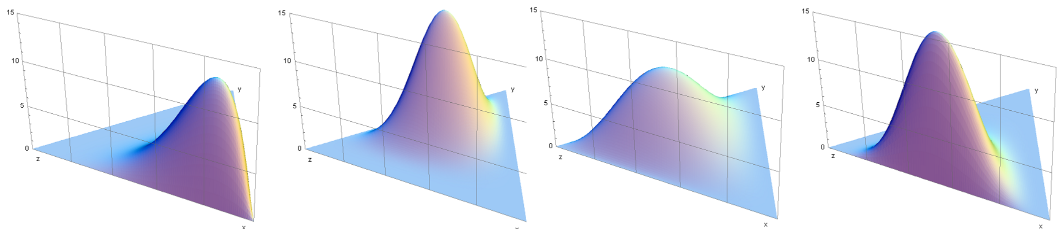
$$\Pi\Big(P \in \mathscr{P} : -P_0 \log dP/dP_0 < \epsilon\Big) > 0$$

*for all $\epsilon > 0$. Then the posterior is consistent at $P_0$ in the Hellinger topology*

# The Dirichlet process

**Definition 7.1** *(Dirichlet distribution)*
*A random variable* $p = (p_1, \ldots, p_k)$ *with* $p_l \geq 0$ *and* $\sum_l p_l = 1$ *is Dirichlet distributed with parameter* $\alpha = (\alpha_1, \ldots, \alpha_k)$, $p \sim D_\alpha$, *if it has density*

$$f_\alpha(p) = C(\alpha) \prod_{l=1}^{k} p_l^{\alpha_l - 1}$$



**Definition 7.2** *(Dirichlet process, Ferguson 1973-74)*
*Let* $\alpha$ *be a finite measure on* $(\mathscr{X}, \mathscr{B})$. *The Dirichlet process* $P \sim D_\alpha$ *is defined by, (for all finite msb partitions* $A = \{A_1, \ldots, A_k\}$ *of* $\mathscr{X}$*)*

$$\big(P(A_1), \ldots, P(A_k)\big) \sim D_{(\alpha(A_1), \ldots, \alpha(A_k))}$$

# Weak consistency with Dirichlet priors

**Theorem 8.1** *(Dirichlet consistency)*
*Let $X_1, X_2, \ldots$ be an i.i.d.-sample from $P_0$ If $\Pi$ is a Dirichlet prior $D_\alpha$ with finite $\alpha$ such that* $\mathrm{supp}(P_0) \subset \mathrm{supp}(\alpha)$*, the posterior is consistent at $P_0$ in the weak model topology*

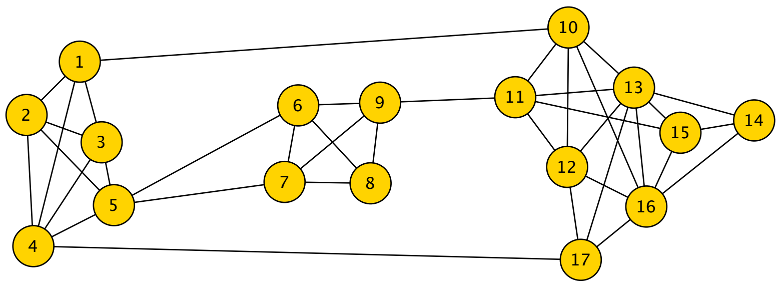
**Remark 8.2** *Priors are not necessarily KL for consistency*

**Remark 8.3** *(Freedman (1965))*
*Dirichlet distributions are tailfree: if $A'$ refines $A$ and $A'_{i1} \cup \ldots \cup A'_{il_i} = A_i$, then $(P(A'_{i1}|A_i), \ldots, P(A'_{il_i}|A_i) : 1 \leq i \leq k)$ is independent of $(P(A_1), \ldots, P(A_k))$.*

**Remark 8.4** $X^n \mapsto \Pi(P(A)|X^n)$ *is $\sigma_n(A)$-measurable where $\sigma_n(A)$ is generated by products of the form $\prod_{i=1}^n B_i$ with $B_i = \{X_i \in A\}$ or $B_i = \{X_i \notin A\}$.*

# Stochastic Block Model

**Definition 9.1** *At step $n$, nodes belong to one of $K_n$ unobserved classes: $\theta_i$. We estimate $\theta = (\theta_1, \ldots, \theta_n) \in \Theta_n$ upon observation of $X^n = \{X_{ij} : 1 \leq 1 < j \leq n\}$. Edges $X_{ij}$ occur independently with probabilities $Q_{ij}(\theta) = Q(\theta_i, \theta_j)$. The (expected) degree is denoted $\lambda_n$.*




An SBM network realisation: $n = 17$, $K_n = 3$, $\lambda_n \approx 4.48$

# Bayesian and Frequentist testability

For $B, V$ be two (disjoint) model subsets

**Definition 10.1** *Uniform (or minimax) testability*

$$\sup_{P\in B} P^n\phi_n \to 0, \qquad \sup_{Q\in V} Q^n(1-\phi_n) \to 0$$

**Definition 10.2** *Pointwise testability for all $P \in B$, $Q \in V$*

$$\phi_n \xrightarrow{P\text{-a.s.}} 0, \qquad \phi_n \xrightarrow{Q\text{-a.s.}} 1$$

**Definition 10.3** *Bayesian testability for Π-almost-all $P \in B$, $Q \in V$*

$$\phi_n \xrightarrow{P\text{-a.s.}} 0, \qquad \phi_n \xrightarrow{Q\text{-a.s.}} 1$$

# Examples of uniform test sequences

**Lemma 11.1** *(Uniform Hellinger tests) Let $B, V \subset \mathscr{P}$ be convex with $H(B,V) > 0$. There exist a $D > 0$ and uniform test sequence $(\phi_n)$ s.t.*

$$\sup_{P \in B} P^n \phi_n \le e^{-nD}, \qquad \sup_{Q \in V} Q^n (1 - \phi_n) \le e^{-nD}$$

**Lemma 11.2** *(Minimax weak tests) Let $n \ge 1$, $\epsilon > 0$, $P_0 \in \mathscr{P}$ and a msb $f : \mathscr{X}^n \to [0,1]$ be given. Define*

$$B = \Big\{ P \in \mathscr{P} \,:\, |(P^n - P_0^n) f| < \epsilon \Big\}, \quad V = \Big\{ P \in \mathscr{P} \,:\, |(P^n - P_0^n) f| \ge 2\epsilon \Big\}$$

*There exist a $D > 0$ and uniform test sequence $(\phi_n)$ s.t.*

$$\sup_{P \in B} P^n \phi_n \le e^{-nD}, \qquad \sup_{Q \in V} Q^n (1 - \phi_n) \le e^{-nD}$$

# Testing in the Stochastic Block Model

Assume there is are $q_n$ s.t. $0 < q_n < Q_{ij} < 1 - q_n < 1$

**Lemma 12.1** *For given, $B_n, V_n \subset \Theta_n$, there exists a test $\phi_n$ s.t.*

$$\max_{\theta \in B_n} P_{\theta,n}\phi_n \le e^{-8q_n(1-q_n)\,a_n^2 + \log \#(V_n)}$$

$$\max_{\theta' \in V_n} P_{\theta',n}(1-\phi_n) \le e^{-8q_n(1-q_n)\,a_n^2 + \log \#(B_n)}$$

*where* $a_n^2 = \inf_{\theta \in B_n} \inf_{\theta \in V_n} \sum_{i<j} (Q_{ij}(\theta) - Q_{ij}(\theta'))^2$

Note: $\log \#(V_n), \log \#(B_n) \le n \log(K_n)$

**Remark 12.2** *Sharper tests are available (Bickel & Chen (2009); Choi, Wolfe & Airoldi (2012); Mossel, Neeman & Sly (2012, 2014); Abbe, Bandeira & Hull (2014))*

# Part II

## Bayesian testability

## and prior-a.s.-consistency

# A posterior concentration inequality

**Lemma 14.1** *Let* $(\mathscr{P},\mathscr{G})$ *be given. For any prior* $\Pi$, *any test function* $\phi$ *and any* $B, V \in \mathscr{G}$,

$$\int_B P\Pi(V|X)\, d\Pi(P) \leq \int_B P\phi\, d\Pi(P) + \int_V Q(1-\phi)\, d\Pi(Q)$$

**Corollary 14.2** *Consequently, in i.i.d.-context, for any sequences* $(\Pi_n)$, $(B_n)$, $(V_n)$ *such that* $B_n \cap V_n = \varnothing$ *and* $\Pi_n(B_n) > 0$, *we have,*

$$\int P^n \Pi(V_n|X^n)\, d\Pi_n(P|B_n) \leq \frac{1}{\Pi(B_n)}\Big(\int_{B_n} P^n\phi_n\, d\Pi_n(P) + \int_{V_n} Q^n(1-\phi_n)\, d\Pi_n(Q)\Big)$$

# Martingale convergence

**Proposition 15.1** *Let $(\mathscr{P}, \mathscr{G}, \Pi)$ be given. For any $B, V \in \mathscr{G}$, the following are equivalent,*

*(i) There exist Bayesian tests $(\phi_n)$ for $B$ versus $V$;*

*(ii) There exist tests $(\phi_n)$ such that,*

$$\int_B P^n \phi_n \, d\Pi(P) + \int_V Q^n (1-\phi_n) \, d\Pi(Q) \to 0,$$

*(iii) For $\Pi$-almost-all $P \in B$, $Q \in V$,*

$$\Pi(V|X^n) \xrightarrow{P\text{-a.s.}} 0, \qquad \Pi(B|X^n) \xrightarrow{Q\text{-a.s.}} 0$$

**Remark 15.2** *Interpretation distinctions between model subsets are Bayesian testable, iff they are picked up by the posterior asymptotically, if(f), the Bayes factor for B versus V is consistent*

# Prior-almost-sure consistency

**Theorem 16.1** *Let Hausdorff* $\mathscr{P}$ *with Borel prior* $\Pi$ *be given. Assume that for* $\Pi$*-almost-all* $P \in \mathscr{P}$ *and any open nbd* $U$ *of* $P$*, there exist a* $B \subset U$ *with* $\Pi(B) > 0$ *and Bayesian tests* $(\phi_n)$ *for* $B$ *versus* $\mathscr{P} \setminus U$*. Then the posterior is consistent at* $\Pi$*-almost-all* $P \in \mathscr{P}$

**Remark 16.2** *Let* $\mathscr{P}$ *be a Polish space and assume that all* $P \mapsto P^n(A)$ *are Borel measurable. Then, for any prior* $\Pi$*, any Borel set* $V \subset \mathscr{P}$ *is Bayesian testable versus* $\mathscr{P} \setminus V$*.*

**Corollary 16.3** *(More than) Doob's 1948 theorem*

# Part III

## Pointwise testability
## and frequentist consistency

# Le Cam's inequality

**Definition 18.1** *For* $B \in \mathscr{G}$ *such that* $\Pi(B) > 0$, *the local prior predictive distribution is* $P_n^{\Pi|B} = \int P^n \, d\Pi(P|B)$.

**Remark 18.2** *(Le Cam, unpublished (197?) and (1986))*
*Rewrite the posterior concentration inequality*

$$P_0^n \Pi(V_n|X^n) \leq \left\| P_0^n - P_n^{\Pi|B_n} \right\| + \int P^n \phi_n \, d\Pi(P|B_n) + \frac{\Pi(V_n)}{\Pi(B_n)} \int Q^n (1 - \phi_n) \, d\Pi(Q|V_n)$$

**Remark 18.3** *For some* $b_n \downarrow 0$, $B_n = \{P \in \mathscr{P} : \|P^n - P_0^n\| \leq b_n\}$,

$$a_n^{-1} \, \Pi(B_n) \to \infty$$

**Remark 18.4** *Useful in parametric models but "a considerable nuisance" [sic] (Le Cam (1986)) in non-parametric context*

# Schwartz's theorem revisited

**Remark 19.1** *Suppose that for all $\delta > 0$, there is a $B$ s.t. $\Pi(B) > 0$ and for all $P \in B$ and large enough $n$*

$$P_0^n \Pi(V|X^n) \leq e^{n\delta} P^n \Pi(V|X^n)$$

*then (by Fatou) for large enough $m$*

$$\sup_{n \geq m} \Big[ (P_0^n - e^{n\delta} P_n^{\Pi|B}) \Pi(V|X^n) \Big] \leq 0$$

**Theorem 19.2** *Let $\mathscr{P}$ be a model with KL-prior $\Pi$; $P_0 \in \mathscr{P}$. Let $B, V \in \mathscr{G}$ be given and assume that $B$ contains a KL-neighbourhood of $P_0$. If there exist Bayesian tests for $B$ versus $V$ of* exponential power *then*

$$\Pi(V|X^n) \xrightarrow{P_0-a.s.} 0$$

**Corollary 19.3** *(Schwartz's theorem)*

# Remote contiguity

**Definition 20.1** *Given* $(P_n)$, $(Q_n)$ *of prob msr's,* $Q_n$ *is contiguous w.r.t.* $P_n$ $(Q_n \lhd P_n)$*, if for any msb* $\psi_n : \mathscr{X}^n \to [0,1]$

$$P_n\psi_n = o(1) \quad \Rightarrow \quad Q_n\psi_n = o(1)$$

**Definition 20.2** *Given* $(P_n)$, $(Q_n)$ *of prob msr's and a* $a_n \downarrow 0$, $Q_n$ *is* $a_n$*-remotely contiguous w.r.t.* $P_n$ $(Q_n \lhd a_n^{-1} P_n)$*, if for any msb* $\psi_n : \mathscr{X}^n \to [0,1]$

$$P_n\psi_n = o(a_n) \quad \Rightarrow \quad Q_n\psi_n = o(1)$$

**Remark 20.3** *Contiguity is stronger than remote contiguity*
*note that* $Q_n \lhd P_n$ *iff* $Q_n \lhd a_n^{-1} P_n$ *for all* $a_n \downarrow 0$.

**Definition 20.4** *Hellinger transform* $\psi(P,Q;\alpha) = \int p^\alpha q^{1-\alpha}\, d\mu$

# Le Cam's first lemma

**Lemma 21.1** *Given* $(P_n)$, $(Q_n)$ *like above,* $Q_n \lhd P_n$ *iff any of the following holds:*

(i) *If* $T_n \xrightarrow{P_n} 0$, *then* $T_n \xrightarrow{Q_n} 0$

(ii) *Given* $\epsilon > 0$, *there is a* $b > 0$ *such that* $Q_n(dQ_n/dP_n > b) < \epsilon$

(iii) *Given* $\epsilon > 0$, *there is a* $c > 0$ *such that* $\|Q_n - Q_n \wedge c\,P_n\| < \epsilon$

(iv) *If* $dP_n/dQ_n \xrightarrow{Q_n\text{-w.}} f$ *along a subsequence, then* $P(f > 0) = 1$

(v) *If* $dQ_n/dP_n \xrightarrow{P_n\text{-w.}} g$ *along a subsequence, then* $Eg = 1$

(vi) $\liminf_n \psi(P_n, Q_n; \alpha) \to 1$ *as* $\alpha \uparrow 1$

# Criteria for remote contiguity

**Lemma 22.1** *Given* $(P_n)$, $(Q_n)$, $a_n \downarrow 0$, $Q_n \lhd a_n^{-1} P_n$ *if any of the following holds:*

*(i)* *For any bnd msb* $T_n : \mathscr{X}^n \to \mathbb{R}$, $a_n^{-1} T_n \xrightarrow{P_n} 0$, *implies* $T_n \xrightarrow{Q_n} 0$

*(ii)* *Given* $\epsilon > 0$, *there is a* $\delta > 0$ *s.t.* $Q_n(dP_n/dQ_n < \delta\, a_n) < \epsilon$ *f.l.e.n.*

*(iii)* *There is a* $b > 0$ *s.t.* $\liminf_{n\to\infty} b\, a_n^{-1}\, P_n(dQ_n/dP_n > b\, a_n^{-1}) = 1$

*(iv)* *Given* $\epsilon > 0$, *there is a* $c > 0$ *such that* $\|Q_n - Q_n \wedge c\, a_n^{-1} P_n\| < \epsilon$

*(v)* *Under* $Q_n$, $(a_n\, dQ_n/dP_n)$ *are r.v.'s and every subseq has a weakly convergent subseq*

*(vi)* $\liminf_n \lim_{\alpha\uparrow 1} a_n^{-\alpha}\, \psi(P_n, Q_n; \alpha) > 0$

# Beyond Schwartz

**Theorem 23.1** *Let* $(\mathscr{P}, \mathscr{G})$ *with priors* $(\Pi_n)$ *and* $(X_1, \ldots, X_n) \sim P_0^n$ *be given. Assume there are* $B, V \in \mathscr{G}$ *with* $\Pi(B) > 0$ *and* $a_n \downarrow 0$ *s.t.*

*(i) There exist Bayesian tests for* $B$ *versus* $V$ *of power* $a_n$*,*

$$\int_B P^n \phi_n \, d\Pi_n(P) + \int_V Q^n (1 - \phi_n) \, d\Pi_n(Q) \leq a_n$$

*(ii) The sequence* $P_0^n$ *satisfies* $P_0^n \lhd a_n^{-1} \, P_n^{\Pi_n|B}$

*Then* $\Pi_n(V|X^n) \xrightarrow{P_0} 0$

# Application to consistency I

**Remark 24.1** *(Schwartz (1965))*
*Take $P_0 \in \mathscr{P}$, and define*

$$V_n = \{P \in \mathscr{P} : H(P, P_0) \geq \epsilon\}$$

$$B_n = \{P : -P_0 \log dP/dP_0 < \epsilon^2\}$$

*With $N(\epsilon, \mathscr{P}, H) < \infty$, and $a_n$ of form $\exp(-nD)$ the theorem proves Hellinger consistency with KL-priors.*

# Application to consistency II

**Remark 25.1** *Dirichlet posteriors* $X^n \mapsto \Pi(P(A)|X^n)$ *are msb* $\sigma_n(A)$ *where* $\sigma_n(A)$ *is generated by products of the form* $\prod_{i=1}^n B_i$ *with* $B_i = \{X_i \in A\}$ *or* $B_i = \{X_i \notin A\}$.

**Remark 25.2** *(Freedman (1965), Ferguson (1973), Lo (1984), ...)* *Take* $P_0 \in \mathscr{P}$, *and define*

$$V_n = V := \{P \in \mathscr{P} : |(P_0 - P)f| \geq 2\epsilon\}$$

$$B_n = B := \{P : |(P_0 - P)f| < \epsilon\}$$

*for some bounded, measurable* $f$. *Impose remote contiguity only for* $\psi_n$ *that are* $\sigma_n(A)$*-measurable! Take* $a_n$ *of form* $\exp(-nD)$. *The theorem then proves weak consistency with a Dirichlet prior* $D_\alpha$, *if* $\mathrm{supp}(P_0) \subset \mathrm{supp}(\alpha)$.

# Consistency with $n$-dependent neighbourhoods

**Theorem 26.1** *Let* $(\mathscr{P}, \mathscr{G})$ *with priors* $(\Pi_n)$ *and* $(X_1, \ldots, X_n) \sim P_0^n$ *be given. Assume there are* $B_n, V_n \in \mathscr{G}$ *and* $a_n, b_n \geq 0$*,* $a_n \downarrow 0$ *s.t.*

*(i) There exist Bayesian tests for* $B_n$ *versus* $V_n$ *of power* $a_n$*,*

$$\int_{B_n} P^n \phi_n \, d\Pi_n(P) + \int_{V_n} Q^n (1 - \phi_n) \, d\Pi_n(Q) \leq a_n$$

*(ii) The prior mass of* $B_n$ *is lower-bounded by* $b_n$*,* $\Pi_n(B_n) \geq b_n$

*(iii) The sequence* $P_0^n$ *satisfies* $P_0^n \lhd b_n a_n^{-1} \, P_n^{\Pi_n | B_n}$

*Then* $\Pi_n(V_n | X^n) \xrightarrow{P_0} 0$

# Application to the posterior rate of convergence

**Remark 27.1** *(Ghosal-Ghosh-vdVaart (2000))*
*Take $P_0 \in \mathscr{P}$, and define*

$$V_n = \{P \in \mathscr{P} : H(P, P_0) \geq \epsilon_n\}$$
$$B_n = \{P : -P_0 \log dP/dP_0 < \epsilon_n^2,\ P_0 \log^2 dP/dP_0 < \epsilon_n^2\}$$

*With $\log N(\epsilon_n, \mathscr{P}, H) \leq n\epsilon_n^2$, and $a_n$ and $b_n$ of form $\exp(-Kn\epsilon_n^2)$ the theorem proves Hellinger consistency at rate $\epsilon_n$ with GGV-priors.*

**Remark 27.2** *Other $B_n$ are possible! (see Kleijn and Zhao (201x))*

# Consistent Bayes factors

**Theorem 28.1** *Let the model* $(\mathscr{P},\mathscr{G})$ *with priors* $(\Pi_n)$ *be given. Given* $B,V\in\mathscr{G}$ *with* $\Pi(B),\Pi(V)>0$ *s.t.*

*(i) There exist Bayesian tests for* $B$ *versus* $V$ *of power* $a_n\downarrow 0$*,*

$$\int_B P^n\phi_n\,d\Pi_n(P)+\int_V Q^n(1-\phi_n)\,d\Pi_n(Q)\le a_n$$

*(ii) For every* $P\in B$*,* $P^n \lhd a_n^{-1}\,P_n^{\Pi_n|B}$

*(iii) For every* $Q\in V$*,* $Q^n \lhd a_n^{-1}\,P_n^{\Pi_n|V}$

*Then the posterior odds or Bayes factors,*

$$B_n=\frac{\Pi(B|X^n)}{\Pi(V|X^n)}\frac{\Pi(V)}{\Pi(B)}$$

*for* $B$ *versus* $V$ *are consistent.*